ABSTRACT OF THE DISCLOSURE

A method and system for efficiently, quickly, and economically buffering data in a network node. Incoming data from the network is received by the network node. This data is first temporarily stored in a tail cache. Blocks of incoming data are stored in the tail cache. When a predetermined number of N blocks of data are stored in the tail cache, a single write operation is initiated to write the N blocks of data from the tail cache to a section of main memory. When a head cache becomes empty, it requests data from the main memory. The predetermined number of N blocks of data from the main memory is transferred to the head cache in a single memory access operation. Eventually, the network node is allowed to transmit data onto the network, whereupon the head cache outputs its data onto the network. In the present invention, the tail cache and the head cache are comprised of relatively small, but fast SRAM memory; whereas the main memory is comprised of slower, but less expensive DRAM memory. By implementing this caching scheme, the super block of N blocks is always filled with data, thereby maintaining full space and bandwidth efficiencies at all times.